ECON 5033 Econometrics I – Lecture 2

Simple Linear Regression Model

Eric S. Lin

Department of Economics, National Tsing Hua University

September 24, 2024

Outline

- Econometrics
- Basic Setup
- Assumptions
- Ordinary Least Squares
- R-square
- Properties of the Estimator
- Sampling Distribution
- Hypothesis Testing
- Prediction

Why is "Econometrics"?

- Why do we need to learn econometrics?
- Simple answer is: it provides necessary tools to an [economist] to derive useful information about [economic] policies (the problems and solutions) using the available data
- Usually there are two steps to explore economic policies or economic issues
    1. First, knowing the theory and establishing the set of hypotheses which is understood by studying economics
    2. Second, once the theory is known, testing the theory or the hypotheses using various techniques. This second step is achieved through studying econometrics
- Some examples apply the two-step procedure

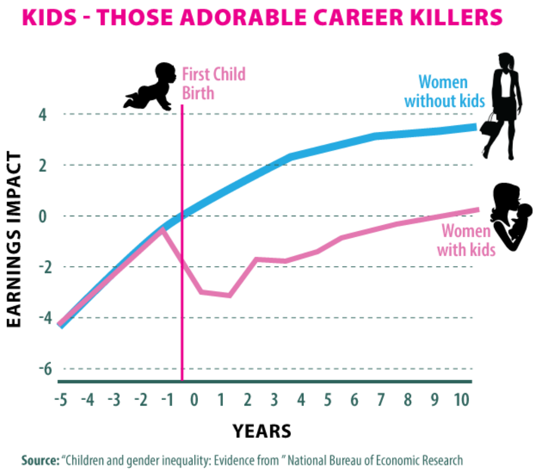
Example [I] Using "Econometrics"

- Consider the wage subsidy and employment (like 20K policy). Follow our previous two-step analysis procedure:
  1. First, an economist would establish a theory that an increase in wage subsidy would increase employment
  2. Second, an econometrician would obtain historical data that is available to test the theory
- It is quite likely that the econometrician would conclude that the theory is not correct or robust
- The increase in wage subsidy would increase youth employment but not non-youth employment. This calls for altering the economic theory
- Hence, studying econometrics is crucial in understanding economic policy issues

Example [II] Using "Econometrics"

- Consider the gender pay gap issue, which is hot in economics and sociology. Follow our previous two-step analysis procedure:
  1. First, the economist can establish various theories by focusing on factors like differences in experience and education. The theory may be that in general, men tend to have higher education and experience which explains why they receive a higher pay than women – too abstract if no further verification
  2. Second, an econometrician would collect historical data on difference in gender pay, educational levels, experience, etc. and use statistical techniques like linear regression analysis to check for the factors that may explain the gender pay gap
- The conclusion of the exercise by an econometrician would be whether a factor like education or experience is important or significant enough in explaining gender pay gap

Example [II] Using "Econometrics"



Example [III] Using "Econometrics"

- Suppose the government has to forecast the rate of unemployment in the country in the coming year. Follow our previous two-step analysis procedure:
  1. First, in economics, there is a very important theory known as Philip's curve. According to this theory, the inflation rate and the unemployment rate are inversely related. This theory is abstract unless tested using data.
  2. Second, econometricians studied historical data of U.S. (on the unemployment rate and all other factors on which unemployment rate may depend) for the period around the 1970's
- It is found that both inflation and unemployment were rising in the U.S. at the same time
- This posed a challenge to the economic theory given by the economist Philips.

Define "Econometrics" Academically

Academic Definition of Econometrics

- Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference – P.A Samuelson et al., (1954) "Report of the Evaluative Committee for Econometrica," *Econometrica*, 22(2).
- Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena – Arthur S. Goldberger, (1964) *Econometric Theory*, John Wiley & Sons, Inc., New York.
  - tools based on computer science are crucial nowadays
  - the most frequently used starting point in modern econometrics: linear regression model

Econometrics: Further Remarks

- Econometrics is much more than just "statistics using economic data," although it is of course very closely related to statistics – it is much broader
- It is important for decision making in regard to governments, businesses, policy organizations, central banks, financial services firms, and economic consulting firms around the world

Econometrics: Government

Example

Governments, central banks and policy organizations use econometric models to guide monetary policy, fiscal policy, as well as education and training, health, law amendment, immigration, and transfer policies. e.g., is wearing masks effective to reduced covid-19 confirmed cases? female leader matters? keeping social distance matters?

Econometrics: Business

Example

Businesses use econometrics for strategic planning tasks. These include management strategy of all types including operations management and control (hiring, production, inventory, investment, new market entry,...), marketing (pricing, distributing, advertising,...), accounting (budgeting revenues and expenditures), and so on.

Econometrics: Financial Services

Example

Econometric models are also crucial in financial services, including asset management, asset pricing, mergers and acquisitions, investment banking, and insurance. Portfolio managers, for example, are keenly interested in the empirical modeling and understanding of asset returns (stocks, bonds, exchange rates, commodity prices, ...).

Econometrics: Consulting Firms

Example

Econometrics is central to the work of a wide variety of consulting firms, many of which support the business functions already mentioned. Litigation support, for example, is also a very active area, in which econometric models are routinely used for damage assessment (e.g., lost earnings), "but for" analyses, and so on, e.g., evaluating counterfactual income for disables by car accident

Econometrics: My Point of View

- Learning econometrics is really worth for different disciplines, not only restricted to economics majors
- Especially in the era of big data, having the knowledge of math & statistics (econometrics), computer science (data structure, data inquiry language), big data analytics (machine learning, deep learning, text/image/audio/video processing), you will be capable of doing many jobs
- With the training from economics profession, you will be equipped with a good smell and thus sense a good issue
- Learning econometrics in economics program will be highly likely to make you unique (conditional on your computer science knowledge)

AJPS

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Can Descriptive Representation Help the Right Win Votes from the Poor? Evidence from Brazil

Zuheir Desai IE University
Anderson Frey University of Rochester

Abstract: *The electoral success of the Right in poor nations is typically attributed to nonpolicy appeals such as clientelism. Candidate profiles are usually ignored because if voters value class-based descriptive representation, it should be the Left that uses it. In this article, we develop and test a novel theory of policy choice and candidate selection that defies this conventional wisdom: it is the Right that capitalizes on descriptive representation in high-poverty areas. The Right is only competitive in poor regions when it matches the Left's pro-poor policies. To credibly shift its position, it nominates candidates who are descriptively closer to the poor. Using a regression discontinuity design in Brazilian municipal elections, we show that Right-wing mayors spend less on the poor than Left-wing mayors only in low-poverty municipalities. In high-poverty municipalities, not only does the Right match the Left's policies, it also does so while nominating less educated candidates.*

Verification Materials: The materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/DQTIR4.

Right-wing parties often win elections in developing nations where voters are overwhelmingly poor. Prevailing explanations for this puzzle typically focus on how they build a portfolio of electoral appeals such as clientelism (Murillo and Calvo 2019), ethnic mobilization (Huber 2017), positioning on "social" dimensions (Tavits and Potter 2015), or private provision of social services (Thachil 2014). The case of Brazil is similar: clientelism and personalistic politician–voter ties have been the primary explanation for why "conservative parties fare best electorally among relatively poor, less educated" voters (Mainwaring, Meneguello, and Power 2000), despite the fact that the Left is more likely to support redistributive policies.

Not surprisingly, these explanations seldom focus on the descriptive profile of the candidates nominated by the Right. The literature on political behavior suggests that voters value descriptive representation (Carnes and Lupu 2016; Dal Bo et al. 2019), and are more likely to trust and feel included by politicians descriptively closer to them (Gay 2002; Hayes and Hibbing 2017; Lawless 2004). In turn, when politicians stress that "I am one of you," their common identity helps them to better understand the needs of voters (Carnes and Lupu 2015), and provides incentives for the betterment of the status of their shared social group (Shayo 2009). Thus, if there are electoral returns to class-based descriptive representation, it is natural to expect that Left-wing parties are the ones that capitalize on it in poor areas. Former Brazilian president Lula (2003–10) is a clear example. He often used his lack of education to emphasize his ability to succeed as a politician, and to implement redistributive policies, mentioning, for example, that "a steelworker without a bachelor's degree created more universities than the PhDs that previously governed the country."[1]

However, in this article we uncover an empirical pattern in Brazilian municipalities that at first defies this conventional wisdom: it is the Right that capitalizes on descriptive representation in the poorest areas. We interpret this finding within the literature on party

Zuheir Desai, IE University, IE Tower, 16.22, 28046 Madrid, Spain (zuheir.desai@ie.edu). Anderson Frey, Department of Political Science, University of Rochester, Harkness Hall, 320B, Rochester, NY 14627 (anderson.frey@rochester.edu).

We thank Tasos Kalandrakis, Alexander Lee, Jack Paine, Umberto Mignozzetti, and all participants at APSA 2019, SPSA 2020, and the Political Economy Workshop at Princeton University for comments and suggestions. All errors are our own.

[1]Tania Monteiro, "Lula diz querer eleger alguém para fazer mais do que fez." *Politica*, June 2009, https://bit.ly/2X0U3rq.

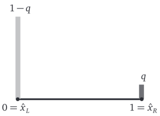
American Journal of Political Science, Vol. 67, No. 3, July 2023, Pp. 671–686

© 2021 The Authors. *American Journal of Political Science* published by Wiley Periodicals LLC on behalf of Midwest Political Science Association.
DOI: 10.1111/ajps.12664

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or

AJPS

FIGURE 1 The Electoral Environment



Note: The light and dark bars represent the distribution of voter types P and A. Poor voters share their ideal point with party L and affluent voters share their ideal point with party R

Model Setup

Our framework is based on Desai (2021), with two ideologically opposed parties L and R. The programmatic Left-Right dimension is defined as the [0,1] interval, and positions closer to 0 represent Leftist, pro-poor policies. The ideal point of party i is given by $\hat{x}_i$. Accordingly, L's ideal point is situated at 0, whereas that of R is situated at 1. There are two classes of voters, poor (P) and affluent (A), which share the ideal points of L and R; respectively. The distribution of voters is indexed by the proportion of A voters, denoted by q. Because we focus on a developing context, we assume that the affluent are always in the minority ($q < \frac{1}{2}$). This setup is reflected in Figure 1.

Parties: Before the election, parties announce policies and choose candidates. The candidate pool for each party contains elite candidates, which are descriptively closer to affluent voters, and nonelite voters, which are descriptively closer to the poor. An observable feature of elitism is, for example, education.

Although policies are chosen from the [0,1] interval, candidate selection is binary. We say that if $c_i = 1$, then i's candidate provides descriptive representation to the group that does not share its ideal point, and $c_i = 0$ otherwise. Let $\mathbf{x}_i = (x_i, c_i) \in [0,1] \times \{0,1\}$ be the policy announcement and candidate choice of party i. Both parties have mixed motivation, that is, they benefit both from holding office and policy outcomes. The weight attached to policy benefits is given by $w \in [0,1]$, and the office benefit is normalized to 1. Denote

$$\varphi(\mathbf{x}_i) = \begin{cases} w\hat{x}_i + (1-w)x_i & \text{if } c_i = 0 \\ x_i & \text{if } c_i = 1 \end{cases} \tag{1}$$

as the final policy implemented by i on winning the election with candidate c_i. The policy function (1) indicates that the more policy-motivated parties are, the more likely it is that their implemented policy deviates from their announced policy position in absence of a descriptively representative candidate. All else equal, both parties face a cost of κ when choosing a nonelite candidate. We interpret this cost framework as context specific. In Supporting Information (SI) Appendix B, we provide an extensive discussion on the motivation behind this assumption. We use both the literature and new empirical evidence to show that this particular cost framework is not only apt to the Brazilian context, but that it also fits the empirical results better than alternative assumptions.[8]

The objective functions of the two parties are given by

$$\begin{aligned} V_L(\mathbf{x}_L, \mathbf{x}_R) = w((1 - F(\mathbf{x}_L, \mathbf{x}_R)) \cdot u_L(\varphi(\mathbf{x}_L)) \\ + F(\mathbf{x}_L, \mathbf{x}_R) \cdot u_L(\varphi(\mathbf{x}_R))) \\ + (1-w)(1-F(\mathbf{x}_L, \mathbf{x}_R)) - (1-c_L)\kappa \end{aligned} \tag{2}$$

$$\begin{aligned} V_R(\mathbf{x}_L, \mathbf{x}_R) = w((1 - F(\mathbf{x}_L, \mathbf{x}_R)) \cdot u_R(\varphi(\mathbf{x}_L)) \\ + F(\mathbf{x}_L, \mathbf{x}_R) \cdot u_R(\varphi(\mathbf{x}_R))) \\ + (1-w)F(\mathbf{x}_L, \mathbf{x}_R) - c_R\kappa, \end{aligned} \tag{3}$$

where $u_i(x) = -|\hat{x}_i - x|$ and $F(\mathbf{x}_L, \mathbf{x}_R)$ is the probability that party R wins the election.

Voter behavior: A voter of class j receives the following utility from party i:

$$u_j(\mathbf{x}_i) = -|\hat{x}_j - \varphi(\mathbf{x}_i)|. \tag{4}$$

Let

$$\Delta u_j(\mathbf{x}_L, \mathbf{x}_R) := u_j(\mathbf{x}_L) - u_j(\mathbf{x}_R) \tag{5}$$

be the utility differential to voter of class j from the candidate–policy pairs of both parties. Each voter j has two idiosyncratic components to her utility, individual and aggregate. The voter has an individual preference η_j for party R, which is drawn identically and independently from a distribution G. This represents how voter j evaluates party R's characteristics on any other criteria other

[8]In summary, on the supply side the candidate pool in Brazil is biased in favor of higher educated candidates due to self-selection into politics. On the demand side, we show that less educated candidates perform worse on many measures of administrative performance indicative of lower valence, and are also worse at brokering votes for their parties in subsequent elections. We also discuss alternative cost structures to the one presented in the text and show that the resulting predictions are at odds from the empirical patterns observed in Brazilian data.

AJPS

than economic policies (e.g., clientelism). In addition to this individual-level idiosyncratic component, all voters receive an aggregate shock ϵ, which is distributed according to the distribution H. This shock represents the aggregate popularity of party L over party R. It affects each voter identically, thereby resulting in parties facing aggregate uncertainty about the election outcome. A negative realization of ϵ means that the electorate is biased toward party R.

Voter j votes for party R if and only if the condition below holds:

$$\begin{aligned} u_j(\mathbf{x}_R) + \eta_j &\geq u_j(\mathbf{x}_L) + \epsilon \\ \iff \eta_j &\geq \Delta u_j(\mathbf{x}_L, \mathbf{x}_R) + \epsilon. \end{aligned}$$

Thus, the proportion of voters voting R is $1 - G(\Delta u_j(\mathbf{x}_L, \mathbf{x}_R) + \epsilon)$. The total vote share for party R is given by the following random variable:

$$\begin{aligned} VS_R(\mathbf{x}_L, \mathbf{x}_R; \epsilon) = &\underbrace{(1-q)(1 - G(\Delta u_P(\mathbf{x}_L, \mathbf{x}_R) + \epsilon))}_{\text{Vote share from poor}} \\ &+ \underbrace{q(1 - G(\Delta u_A(\mathbf{x}_L, \mathbf{x}_R) + \epsilon))}_{\text{Vote share from affluent}}, \quad (6) \end{aligned}$$

and the vote share of party L is analogously $1 - VS_R(\mathbf{x}_L, \mathbf{x}_R; \epsilon)$. Note that the model implies that the smaller Δu_j, the *less* voters vote on the basis of their economic preferences. The probability that R wins the election is the probability that its vote share is greater than that of party L, and is given by

$$F(\mathbf{x}_L, \mathbf{x}_R) := \int \mathbb{I}\left\{VS_R(\mathbf{x}_L, \mathbf{x}_R; \epsilon) \geq \frac{1}{2}\right\} h(\epsilon)\, d\epsilon. \quad (7)$$

The probability that L wins the election is simply $1 - F(\mathbf{x}_L, \mathbf{x}_R)$. We assume that G is uniform on $[-2, 2]$ and H is uniform on $[-\psi, \psi]$, where $\psi < 1$, for tractability.

The game proceeds as follows:

1. Parties choose their policy announcement x_i and candidate c_i.
2. Individual and aggregate shocks η_j and ϵ are realized.
3. Voters sincerely vote for their preferred party.
4. The winning party implements its policy according to $\varphi(\mathbf{x}_i)$.

In what follows, we make the following assumption on the nomination cost.

Assumption 1. *The cost to nominate a nonelite candidate is such that* $\kappa < \frac{\psi}{8}$.

Note that although policy choice is continuous, we adopt a discrete candidate selection framework for the sake of a cleaner exposition. In this class of models, candidate selection follows a cost–benefit analysis. Evidently, descriptive representation as a costly strategy is only viable if the cost is outweighed by the benefit, which is a weighted average of office and policy related benefits. Assumption 1 precludes the existence of a trivial equilibrium where a purely policy-motivated Right rationally chooses to lose the election *for sure* in very poor districts.

Before proceeding, we make a remark on the structure of the model. Our model captures multiple competitive frameworks. The parameter w measures how "programmatic" competition is: It simultaneously measures how motivated parties are on policy, as well as how final policy reflects party ideal points. If $w = 1$, we are in a purely policy-motivated setting where parties care *exclusively* about policy, and no policy position other than the party's ideal point is ex ante credible. On the contrary, if $w = 0$, then parties are purely office motivated, and because party ideal points have no meaning, there is no disconnect between implemented policy and party positions.

Model Results

We present our results in two propositions, each focusing on a particular kind of competitive framework. Party incentives depend on the nature of competition as well as poverty. These two considerations drive parties to reduce or increase programmatic differentiation, which in turn shapes candidate selection patterns.

First, we look at a relatively office motivated competitive framework.

Proposition 1 (Office-motivated framework)**.** *There exists a* $\underline{w} \in (0, 1)$ *such that for all* $w \leq \underline{w}$ *and for all* $q \in (0, \frac{1}{2})$:

1. *party L never nominates a nonelite candidate and implements* $x_L^* = 0$;
2. *party R never nominates a nonelite candidate and implements* $\varphi(\bar{x}_R, 0)$ *upon winning the election, where*

$$\bar{x}_R = \max\left\{\min\left\{\frac{w\psi - (1-2q)(w^2 + (1-w)^2)}{2w(1-w)(1-2q)}, 1\right\}, 0\right\}.$$

It is best to first focus on the case when both parties are purely office-motivated to understand this result. In this case, L and R both choose policies to maximize their probability of winning. Because their policy preferences are irrelevant, all promises are credible. In such a scenario, L and R *always* converge to the policy of the median voter. Because descriptive representation as a tool to establish credibility is unnecessary, both parties nominate elite candidates. When w is positive but small, this logic continues to hold. Although parties have some

AJPS

TABLE 1 Mayor's Partisanship, Education, and Pro-Poor Spending

	Pro-Poor Spending as Percentage of Budget			Education Gap (Winner minus Loser)		
	(1)	(2)	(3)	(4)	(5)	(6)
High poverty	1.062	0.734	0.349	−0.805*	−0.757*	−0.681*
	(0.828)	(0.739)	(0.679)	(0.335)	(0.290)	(0.260)
Pretreatment baseline	59.659	59.630	59.579	−0.033	−0.021	−0.003
Low poverty	−1.904*	−2.030*	−1.928*	0.514†	0.244	0.099
	(0.857)	(0.762)	(0.694)	(0.281)	(0.249)	(0.226)
Pretreatment baseline	50.307	50.352	50.421	0.164	0.136	0.117
Bandwidth	3.97	5.29	6.61	4.05	5.40	6.76
Observations	1544	2026	2464	1566	2061	2504
Bandwidth rules	0.75 × op.	Optimal	1.25 × op.	0.75 × op.	Optimal	1.25 × op.

Note: Standard errors are clustered by municipality (parentheses). The estimates represent the difference in outcomes between municipalities with Right- and Left-wing mayors for each subsample, at the discontinuity. The coefficients come from the estimation of equation (8). †p< .1; *p < .05.

likely to be biased by unobserved municipal characteristics that either influence policies or are correlated with the education of the candidates who run and win elections. We address this problem with an RDD that compares only municipalities where a Right-wing party won (or lost) to a Left-wing party by a close margin.

For the policy variable, the RDD estimates represent the local treatment effect of electing a Right-wing mayor, precisely identified for a municipality were the margin of victory in the election was zero. However, our estimates for the education outcome cannot be interpreted as an *effect* of electing a Rightist politician, given that the nominations happen before elections. Instead, they should be interpreted as the correlation between education and the winner's party ideology.

Nevertheless, there are benefits from also using the RDD to estimate this correlation. First, using an empirical approach consistent with the one used to identify the policy treatment effect, and a comparable sample, the RDD allows to precisely connect both results, as required by our theory. Second, the RDD is a very transparent way to show that this empirical pattern is not driven by a potential correlation between ideology and other observed variables, including other characteristics of candidates. Accordingly, SI Table E.1 shows the balance around the discontinuity of predetermined or fixed covariates. In addition, we also show that the observed relationship between partisanship, education and poverty in Brazil is robust to alternative empirical approaches, such as OLS (cross-section) estimation and panel analysis, and not driven by the RDD assumptions. SI Figure E.1 and Table E.9 show the results of these empirical strategies.

We provide estimates for two subsamples with municipalities with poverty rate above and below the median.[23] Municipal poverty is measured by the share of poor families, estimated by the Ministry of Social Development (MDS).[24] The main estimating equation is

$$\begin{aligned} y_{mt} = {} & \beta_0 + \beta_1 R_{mt} + \beta_2 W_{mt} + \beta_3 R_{mt} W_{mt} \\ & + (\beta_4 + \beta_5 R_{mt} + \beta_6 W_{mt} + \beta_7 R_{mt} W_{mt}) M_{mt} \\ & + \delta_t + \theta_{mt} + \xi_{mt}, \end{aligned} \tag{8}$$

where outcome y_{mt} for municipality m in period t is regressed on the Right-wing dummy R_{mt}, and on the dummy that indicates whether the municipality is in the low-poverty group (W_{mt}). The margin of victory is the difference in the vote share between the winner and runner-up (M_{mt}), δ_t are election fixed-effects, and θ_{mt}

[23]SI Table E.6 shows that the results are robust to the choice of poverty cutoff. In SI Table E.12, we show that they are also robust to nonbinary measures of poverty, and a different definition of the poverty variable that uses the municipal Human Development Index.

[24]This is the base for several federal government benefits including *Bolsa Família*.

RIHE

Research in Higher Education
https://doi.org/10.1007/s11162-024-09795-6

Evaluating the Short-term Causal Effect of Early Alert on Student Performance

Andre Rossi de Oliveira[1]

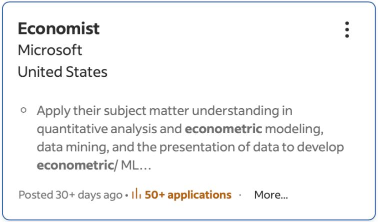
Received: 10 September 2022 / Accepted: 30 April 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract
A little less than half of the students of higher Ed institutions in the US graduate in four years, and only around 60% finish in six years. Retention rates are also less than ideal. Colleges have been experimenting with a variety of programs and policies to address this issue, especially less selective institutions whose rates are significantly lower. In this paper, we evaluate a student success and retention program called Early Alert that was implemented at a public state university in the US with a medium-to-large student body. Our dataset contains several years' worth of information on students' socio-demographic characteristics, class standing and average grades (GPAs), as well as their midterm and final grades in undergraduate courses. We employ several causal inference techniques developed for observational studies and elicit negative average treatment effects on the treated (ATT). Since it is conceivable that unobserved confounders are the real drivers of our empirical results, not the treatment, we carry out two different types of sensitivity analyses. Together with our treatment effect estimations, they lead us to the main conclusion that Early Alert does not improve student performance, at least not in the short run (as measured by course performance), and likely has a negligible impact.

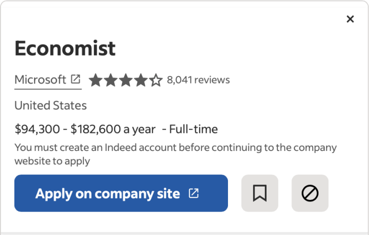
Keywords Causal inference · Early alert · Student success · Matching · Regression

JEL Classification C21 · C55 · I20 · I23

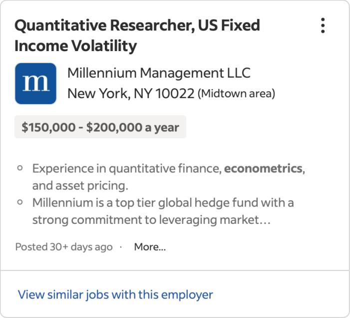
Why Learning Econometrics? From indeed.com



Why Learning Econometrics?



Why Learning Econometrics?



Simple Linear Regression Model

- Economists often use linear regression to quantify a relationship between economic variables.
- A linear regression model between Y and X is of the form:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, ..., n \tag{1}$$

  - Y_i: dependent variable (regressand, outcome)
  - X_i: independent variable (regressor, explanatory variable, covariate, characteristics)
  - ε_i: the error term which may affect Y but unknown to the researcher.
  - (α, β): (unknown, true) parameters to be estimated
    - "parametric" approach

Simple vs. Multiple Regression Models

Example

Wage equation:

$$\ln \text{wage}_i = \alpha + \beta \text{edu}_i + \varepsilon_i.$$

Example

Wage equation:

$$\ln \text{wage}_i = \alpha + \beta \text{edu}_i + \gamma \text{gender}_i + \delta \text{major}_i + \tau \text{industry}_i + ... + \varepsilon_i.$$

Example

Admission method:

$$\text{GPA}_i = \alpha + \beta \text{Application}_i + \gamma \text{FamilyIncome}_i + \delta \text{Club}_i + ... + \varepsilon_i.$$

Linear Predictor

- Suppose (X, Y) are jointly distributed random variables and we wish to "predict" Y from X.
- Our predictor of Y given X is of course some function of X, say $\eta(X)$.
- We will measure the accuracy of prediction by so-called Mean Squared Prediction Error:

$$MSPE = \mathrm{E}[(Y - \eta(X))^2]$$

- One can in fact find the best predictor (in the sense of minimizing MSPE) over all predictors, i.e., all functions of X, namely

$$\eta^*(X) = \mathrm{E}[Y|X = x] = \int y f_{Y|X}(y|x) \mathrm{d}y$$

Best Linear Predictor of Y

- How to obtain the following?

$$\eta^*(X) = \mathrm{E}[Y|X = x] = \int y f_{Y|X}(y|x)\mathrm{d}y$$

Proof.

Let $Y = m(X) + \varepsilon$ where $m(X) = \mathrm{E}[Y|X]$

$$\begin{aligned}
\mathrm{E}[(Y - \eta(X))^2] &= \mathrm{E}[(m(X) + \varepsilon - \eta(X))^2] \\
&= \mathrm{E}[\varepsilon^2] + 2\mathrm{E}[\varepsilon(m(X) - \eta(X))] + \mathrm{E}[(m(X) - \eta(X))^2] \\
&= \mathrm{E}[\varepsilon^2] + \mathrm{E}[(m(X) - \eta(X))^2] \\
&\geq \mathrm{E}[\varepsilon^2]
\end{aligned}$$

$\square$

Best Linear Predictor of Y

- Note that

$$
\begin{aligned}
& \mathrm{E}[\varepsilon \times m(X)] \\
& = \mathrm{E}_X \mathrm{E}[\varepsilon \times m(X) | X] \\
& = \mathrm{E}_X m(X) \mathrm{E}[\varepsilon | X] \\
& = \mathrm{E}_X m(X) \times 0 \\
& = 0
\end{aligned}
$$

Best Linear Predictor of Y

- However, it is sometimes desirable to restrict the class of predictors to so-called linear predictors, which are predictors of the form:

$$\eta^*(X) = Xb = \eta^L(X)$$

- If $\eta^L(X)$ is linear in X, say $Xb = 1 \times b_1 + X_2 \times b_2 + ...X_k \times b_k$
- One can show that

$$\beta = \arg\min_b \mathrm{E}[(Y - Xb)^2] \tag{2}$$

- β minimizes the square loss function:

$$\beta = [\mathrm{E}(X'X)^{-1}]\mathrm{E}(X'Y)$$

- What is the best linear predictor (BLP) of Y?

Population Parameter – An Alternative Derivation

- Go back to regression setup and assume $\mathrm{E}[X_i\varepsilon_i] = 0$ for now
- The regression model in matrix form is:

$$Y = X\beta + \varepsilon \text{ with } \mathrm{E}[X'\varepsilon] = 0$$

- Thus, the parameter β satisfies:

$$\mathrm{E}[X'(Y - X\beta)] = 0 \tag{3}$$

- Solving for the population parameter β gives:

$$\beta = [\mathrm{E}(X'X)^{-1}]\mathrm{E}(X'Y) \;\longleftarrow\; \text{BLP?} \tag{4}$$

Key Assumptions

- Here we list the assumptions in a typical simple linear regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

1. $\mathrm{E}[Y|X_i] = \alpha + \beta X_i$ [linear specification; no model mis-specification]
2. $\mathrm{E}[\varepsilon_i|X_i] = 0$ [no endogeneity]
3. $\mathrm{E}\left[\varepsilon_i^2|X_i\right] = \sigma^2$ [no heteroskedasticity]
4. $\mathrm{E}[\varepsilon_i\varepsilon_j|X_i] = 0$ for all $i \neq j$ [no serial correlation]

Ordinary Least Squares Estimation

- How to implement $\min_b \mathrm{E}[(Y - Xb)^2]$?
- Analogy principle
- Our objective to find α and β which minimize the following criterion function (residual sum of square)

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \alpha - \beta X_i\right]^2 .$$

- FOCs (normal equations) are

$$\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \alpha - \beta X_i\right]^2}{\partial \alpha} = \frac{-2}{n}\sum_{i=1}^{n}\left[Y_i - \alpha - \beta X_i\right] = 0$$

$$\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \alpha - \beta X_i\right]^2}{\partial \beta} = \frac{-2}{n}\sum_{i=1}^{n}X_i\left[Y_i - \alpha - \beta X_i\right] = 0.$$

Ordinary Least Squares Estimation

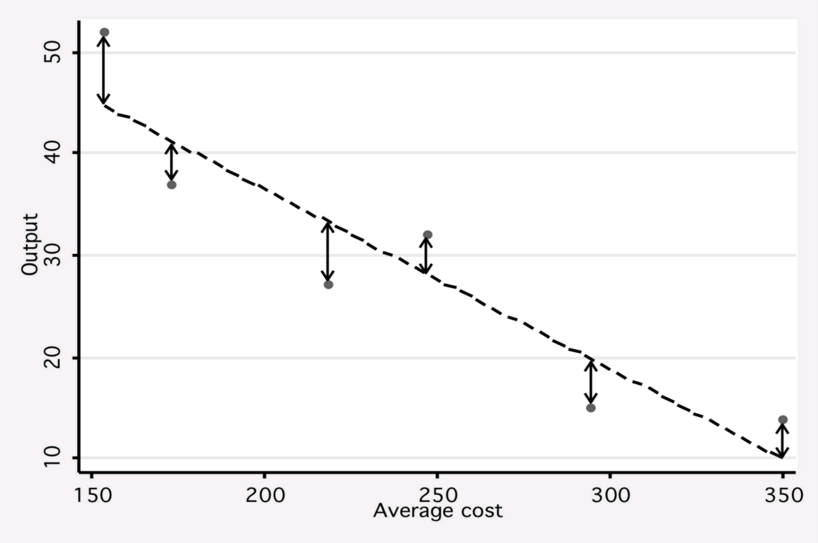
- The solutions for α and β are

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

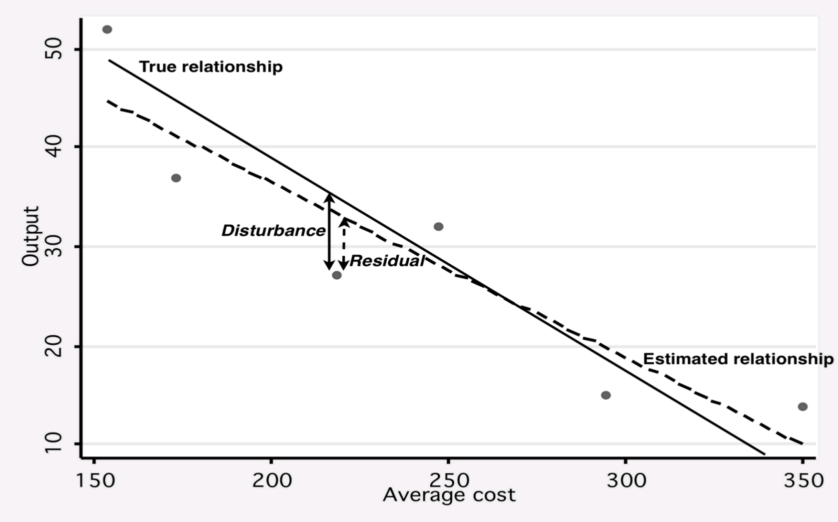
$$\hat{\beta} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = r_{xy}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

- Denote the fitted value (or predicted value) as $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$
- Denote the residual as $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$
- Properties
  - $\sum_{i=1}^{n} e_i = 0$
  - $\sum_{i=1}^{n} X_i e_i = 0$
  - analogy principle again?

OLS – Minimization in Graph



OLS – Residual and Error Term in Graph



OLS as Weighted Average of Y

- We could view $\hat{\beta}$ as a weighted average of Y.
- Actually,

$$\hat{\beta} = \sum_{i=1}^{n} w_i Y_i,$$

where the weight is:

$$w_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

- Note $\sum_{i=1}^{n} w_i = 0$, $\sum_{i=1}^{n} w_i X_i = 1$ and $\sum_{i=1}^{n} w_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

OLS Example

Example

A subsample of the US National Longitudinal Survey (NLS) in 1987 consists of a sample of 3,294 young working individuals, of which 1,569 are female. The average hourly wage rate in this sample equals \$6.31 for males and \$5.15 for females. We try to approximate wages by a linear combination of a constant and a gender dummy ($D_i^m = 1$ for male and 0 for female). The OLS regression result is:

$$\begin{aligned}\hat{Y}_i &= 5.15 + 1.16 \times D_i^m \\ &= \hat{\alpha} + \hat{\beta} \times D_i^m,\end{aligned}$$

where $\hat{\alpha} = \bar{Y}^f$ and $\hat{\beta} = \bar{Y}^m - \bar{Y}^f$.

Goodness of Fit of a Regression

- How to determine the goodness of fit of a regression?
- First decompose:

$$\begin{aligned}\sum_{i=1}^{n}[Y_i - \hat{\alpha} - \hat{\beta}X_i]^2 &= \sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}\left(X_i - \bar{X}\right)]^2 \\ &= \sum_{i=1}^{n}\left[Y_i - \bar{Y}\right]^2 - \hat{\beta}^2 \sum_{i=1}^{n}\left[X_i - \bar{X}\right]^2.\end{aligned}$$

- Now we have:

$$\sum_{i=1}^{n}\left[Y_i - \bar{Y}\right]^2 = \sum_{i=1}^{n} e_i^2 + \hat{\beta}^2 \sum_{i=1}^{n}\left[X_i - \bar{X}\right]^2$$

$$\text{TSS} = \text{RSS} + \text{ESS}$$

R-square

- Rearranging terms leads to the definition of R-square:

$$r^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{ESS}}{\text{TSS}}. \tag{5}$$

- r^2 is a measure of how much of the variation in Y (TSS) is explained by variation in X (ESS)
- r^2 is also the square of the sample correlation coefficient between X and Y in this case
- We will use R^2 to denote the coefficient of determination later on instead of r^2

Small Sample Properties of OLS Estimators

- Small sample properties of $\hat{\alpha}$ and $\hat{\beta}$
- $\hat{\alpha}$ and $\hat{\beta}$ are both unbiased estimators:

$$\mathsf{E}[\hat{\alpha}|X] = \alpha \Rightarrow \mathsf{E}[\hat{\alpha}] = \alpha$$
$$\mathsf{E}[\hat{\beta}|X] = \beta \Rightarrow \mathsf{E}[\hat{\beta}] = \beta.$$

- How about the variances? See [JD]-Appendix 1.2.

$$\text{var}[\hat{\alpha}] = \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}[X_i - \bar{X}]^2}]$$
$$\text{var}[\hat{\beta}] = \frac{\sigma^2}{\sum_{i=1}^{n}[X_i - \bar{X}]^2}.$$

Small Sample Properties of OLS Estimators

- Recall that:

$$\mathrm{var}[\hat{\alpha}] = \sigma^2 [\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} [X_i - \bar{X}]^2}]$$
$$\mathrm{var}[\hat{\beta}] = \frac{\sigma^2}{\sum_{i=1}^{n} [X_i - \bar{X}]^2}.$$

- More sample size and more variation in x will decrease the variance.
- Note that the covariance between $\hat{\alpha}$ and $\hat{\beta}$ is in general not zero unless $\bar{X} = 0$. Derivation see [JD]-Appendix 1.3.

$$\mathrm{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{X}}{\sum_{i=1}^{n} [X_i - \bar{X}]^2}.$$

Gauss-Markov Theorem

Theorem

(Gauss-Markov) The OLS estimator $\hat{\alpha}$ and $\hat{\beta}$ are the Best Linear Unbiased Estimators (BLUE) of α and β.

Proof.

Define an arbitrary unbiased estimator of β, say $b = \sum_{i=1}^{n} q_i Y_i$, such that $\sum_{i=1}^{n} q_i = 0$ and $\sum_{i=1}^{n} q_i X_i = 1$. Variance of b is $\text{var}[b] = \sigma^2 \sum_{i=1}^{n} q_i^2$. Let $q_i = w_i + (q_i - w_i)$. Note that $\sum_{i=1}^{n} q_i^2 = \sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} (q_i - w_i)^2$. Here we use the fact that $\sum_{i=1}^{n} w_i (q_i - w_i) = 0$. Therefore, we end up with $\text{var}[b] = \text{var}[\hat{\beta}] + \sigma^2 \sum_{i=1}^{n} (q_i - w_i)^2$ and the result follows. $\square$

- Limitations of the Gauss-Markov Theorem?

Exact Sampling Distributions

- For simplicity, we assume normality to derive the sampling distributions [Assumption #5]

$$\varepsilon_i | X \sim \mathcal{N}\left(0, \sigma^2\right)$$

- We have the exact (conditional) sampling distribution of the coefficient estimators.

$$\hat{\beta} \sim \mathcal{N}(\beta, \text{var}[\hat{\beta}]) = \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}[X_i - \bar{X}]^2}\right)$$

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \text{var}\left[\hat{\alpha}\right]\right) = \mathcal{N}\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}[X_i - \bar{X}]^2}\right]\right).$$

Estimation of Variance

- To do hypothesis testing one needs to estimate $\text{var}[\varepsilon_i|X]$.
- If we happen to know ε_i, it is natural to estimate it by

$$\frac{1}{n-1}\sum_{i=1}^{n}[\varepsilon_i-\bar{\varepsilon}]^2$$

- However, ε_i is unobservable $\rightsquigarrow$ e_i is useful.
- Estimator of σ^2 is

$$s^2=\frac{1}{n-2}\sum_{i=1}^{n}[e_i-\bar{e}]^2=\frac{1}{n-2}\sum_{i=1}^{n}e_i^2$$

- s^2 is indeed an unbiased estimator of σ^2. We will show that

$$\frac{(n-2)\,s^2}{\sigma^2}\sim\mathcal{X}^2\,(n-2)\,.$$

t-test

- Once we establish the sampling distribution of the coefficient estimators, it's time to conduct hypothesis testing.
- The most general type of test is

$$H_o : \beta = \beta_o$$
$$H_1 : \beta \neq \beta_o$$

- A *t*-test is applicable instead of a z-test
- Let's form a *t*-statistic:

$$t_\beta = \frac{\hat{\beta} - \beta_o}{\sqrt{s^2 / \sum_{i=1}^n \left[X_i - \bar{X}\right]^2}}.$$

t-test

- Does t_β obey a t-distribution?
- Write

$$t_\beta = \left[\frac{\hat{\beta} - \beta_o}{\sqrt{\sigma^2 / \sum_{i=1}^n \left[X_i - \bar{X} \right]^2}} \right] / \sqrt{\frac{(n-2)\, s^2}{\sigma^2} / (n-2)}$$
$$= \frac{\mathcal{N}(0,1)}{\sqrt{\mathcal{X}^2 (n-2) / (n-2)}}$$

- One can show $\mathcal{N}(0,1) \perp \sqrt{\mathcal{X}^2 (n-2) / (n-2)}$.
- Thus, $t_\beta \sim t(n-2)$

Confidence Interval

- Decision rule is

$$\text{Reject } H_o \text{ if } |t_\beta| > t\,(n-2, \gamma/2)\,,$$

where γ is the nominal size or significance level. That is

$$\Pr\left[-t\,(n-2,\gamma/2) < t_\beta < t\,(n-2,\gamma/2)\right] = 1-\gamma.$$

- Let $s.e. = \sqrt{s^2 / \sum_{i=1}^{n} \left[X_i - \bar{X}\right]^2}$
- The $100(1-\gamma)\,\%$ confidence interval for β is:

$$\hat{\beta} - t\,(n-2,\gamma/2) \times (s.e.) < \beta < \hat{\beta} + t\,(n-2,\gamma/2) \times (s.e.)$$

Analysis of Variances

- Recall that $TSS = ESS + RSS$.
- See the following table:

Variation	Sum of squares	D.F.	Mean Squares
Residual	$\text{RSS} = \sum_{i=1}^{n} e_i^2$	$n-2$	$\text{RSS}/(n-2)$
Regressor	$\text{ESS} = \hat{\beta}^2 \sum_{i=1}^{n} \left[X_i - \bar{X}\right]^2$	1	$\text{ESS}/1$
Total	$\text{TSS} = \sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2$	$n-1$	$\text{TSS}/(n-1)$

Analysis of Variances

- Degrees of freedom: the number of values that can be set arbitrarily.

Sum of squares	D.F.	Restriction
$\text{RSS} = \sum_{i=1}^{n} e_i^2$	$n-2$	$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} X_i e_i = 0$
$\text{ESS} = \hat{\beta}^2 \sum_{i=1}^{n} \left[X_i - \bar{X}\right]^2$	1	only one explanatory variable
$\text{TSS} = \sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2$	$n-1$	$\sum_{i=1}^{n} \left[Y_i - \bar{Y}\right] = 0$

Test for "Existence" of Regression

Example

If one would do the test for the "existence" of regression, i.e.

$$H_o : \beta = 0$$
$$H_1 : \beta \neq 0.$$

The test statistic will be

$$F = \frac{ESS/1}{RSS/(n-2)} \sim F(1, n-2).$$

- Note that $F(1, n-2) = [t(n-2)]^2$.

Point Prediction

- Sometimes it is useful to predict the value of Y based on X_o, where X_o may be outside the sample observation.
- Point prediction is

$$\hat{Y}_o = \hat{\alpha} + \hat{\beta} X_o = \bar{Y} + \hat{\beta}\left(X_o - \bar{X}\right).$$

- Ture value Y_o is:

$$Y_o = \alpha + \beta X_o + \epsilon_o. \tag{6}$$

- The average of Y is:

$$\bar{Y} = \alpha + \beta \bar{X} + \bar{\epsilon}. \tag{7}$$

Prediction Error

- Subtracting (6) from (7) gives

$$Y_o = \bar{Y} + \beta \left(X_o - \bar{X}\right) + \epsilon_o - \bar{\epsilon}.$$

- The prediction error is defined as

$$e_o = Y_o - \hat{Y}_o = -(\hat{\beta} - \beta) \left(X_o - \bar{X}\right) + \epsilon_o - \bar{\epsilon}.$$

- The sampling distribution of the prediction error will be:

$$e_o = Y_o - \hat{Y}_o \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(X_o - \bar{X}\right)^2}{\sum_{i=1}^{n} \left[X_i - \bar{X}\right]^2}\right)\right).$$

Test for Prediction

- The t-test becomes

$$\frac{Y_o - \hat{Y}_o}{s\sqrt{1 + \frac{1}{n} + \frac{\left(X_o - \bar{X}\right)^2}{\sum_{i=1}^{n}\left[X_i - \bar{X}\right]^2}}} \sim t\left(n-2\right).$$

- Confidence interval could be formed by

$$\hat{Y}_o \pm t\left(n-2, \gamma/2\right) s\sqrt{1 + \frac{1}{n} + \frac{\left(X_o - \bar{X}\right)^2}{\sum_{i=1}^{n}\left[X_i - \bar{X}\right]^2}}.$$

Mean Prediction

- One may be interested in predicting mean value of Y_o, i.e., $\mathrm{E}[Y|X_o] = \alpha + \beta X_o$.
- In this case, we have

$$\mathrm{E}\left[Y|X_o\right] - \hat{Y}_o = -(\hat{\beta} - \beta)\left(X_o - \bar{X}\right) - \bar{\epsilon}.$$

$$\mathrm{E}\left[Y|X_o\right] - \hat{Y}_o \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n} + \frac{\left(X_o - \bar{X}\right)^2}{\sum_{i=1}^{n}\left[X_i - \bar{X}\right]^2}\right)\right).$$

- t-test and confidence interval are computed similarly.